基于开源关系数据库的 CMS 高可用容灾设计

摘　要：本文叙述了基于 Mysql 全备份和 binlog 日志互补方式实现高负荷新闻发布系统自动化容灾备份的方法，在机房发生全局性故障的情况下，在最短时间内启用容灾机房备份系统进行后续发稿的业务要求。

关键词：内容发布系统；容灾；Mysql；binlog；ElasticSearch
中图分类号：G210.7　　文献标识码：A
文章编号：1671-0134（2018）01-079-03　　DOI：10.19483/j.cnki.11-4653/n.2018.01.029

文 / 赵强　王文峥

引言

中新网 CMS 目前每日上线用户数约 280 人，日发稿万条。每日新增数据库记录 10 万笔以上，高峰期数据库查询达到 500 QPS。由于中新网实行 7×24 小时不间断发稿，一方面系统必须具备容灾能力，主节点发生故障时需要能够切换到备用节点继续提供服务，另一方面要充分考虑备份产生的流量对带宽的占用，以免在业务高峰期时影响业务正常开展。

1.CMS 系统介绍

1. 1 应用简介

CMS 系统是一套 B/S 结构的 Web 应用系统，以 Apache 作为 Web 服务器，后端采用 Mysql 开源数据库作为核心的内容存储。编辑人员在发布系统中上传图文视频等各类稿件素材，经过排版编辑，设定各类栏目及选项，通过发布系统组织成为带有格式的新闻页面，经过多级审核后，将新闻页面通过分发服务器生成静态页面分发至各浏览服务器中提供用户访问。

1. 2 相关技术

1.2.1 MyISAM 数据库引擎

Mysql 数据库随着功能的日益完善和可靠性的不断提高，已经成为互联网平台上应用最广泛的开源关系数据库软件。

Mysql 使用的存储引擎之一 MyISAM，相比其它存储引擎，具备以下优势：文件结构相对独立，数据文件可被压缩，节省存储空间；跨平台可移植性比较好，在备份和恢复时可单独针对某个表进行操作，方便实施以文件为基础的备份策略；查询速度较快；当然，这种引擎也存在一定不足，包括：不支持事务；数据库写入时为表锁定，不支持行级锁定。

由于中新网 CMS 系统要求较高的查询速度，且针对写入时表锁定的情况进行了分表处理，并对数据库查询、写入已经做了大量优化工作。为提高整体运行效率，提升可维护性、可管理型，我们选择使用 MyISAM 引擎作为 Mysql 的存储引擎。

1.2.2 rsync

rsync 全称 remote sync，是一种高效的远程数据同步工具，rsync 使用“rsync 算法”来使本地和远程文件达到同步，这个算法只传送两个文件的不同部分，而不是每次都整份传送，因此传送效率很高，而且可以通过 ssh 方式传输文件，安全性较好。

1.2.3 ElasticSearch

ElasticSearch 是一个基于 Lucene 的搜索服务器，它不仅仅是全文搜索引擎，同时也是优秀的分布式实时文件存储，ES 的集群部署方式可以扩展到上百台服务器，处理 PB 级结构化或者非结构化数据。

2. 传统的备份方案的不足

由于业务数据及结构原因，中新网 CMS 不能简单的使用 Mysql 的主从机制进行数据库备份系统的同步，所以数据库数据的同步依靠传输数据库文件来实现，但同时带来的问题是，庞大的数据库表中，我们只插入或者修改一小部分记录的数据，从数据库文件的角度上看是这个表文件发生了变化，文件级同步则系统对这个很大的数据库文件做整体的差异分析及同步，虽然 rsync 技术可以只同步文件差异部分的内容，但从实际效果看，rsync 对于分析 mysql 自行组织的二进制文件的分析并不精准，往往对一个表的轻微修改会带来很大的传输量。同步数据一方面大量占用了网络的带宽资源，另一方面由于传输时间比较长，加之 Mysql 实例并不能将实时的内容刷新至磁盘文件，所以造成了备用系统与主库依然存在较大差异。采用这种备份方式，会导致业务中断时间很长，在发稿密度极高且 24 小时要求不中断的情况下，这种备份方式是无法满足业务需求的。

3.CMS 高可用系统的设计与实现

3. 1 系统高可用的业务指标

通常信息系统高可用能力用 2 个指标来衡量，包括 RTO（Recovery Time Object） 和 RPO(Recovery Point Object). 指标的定义为：（1）RTO（恢复时间目标），指定故障发生后，从业务停顿到系统恢复可以支持业务运

作时两点之间的时间段。

（2）RPO（恢复点目标），是指一个过去的时间点，当灾难或紧急事件发生时，数据可以恢复到的时间点。

3.2 采用冗余、备份技术进行数据容灾

冗余技术是利用系统的并联模型来提高系统可用性的非常有效的手段 ,CMS 系统的高可用结构也是基于冗余的思路设计实现，希望能够实现两个层级的冗余，第一层级为主机系统的冗余，第二层级为数据中心的冗余。

在 Web 系统中使用冗余结构的关键点在于数据的同步，数据同步的能力决定了 RPO（恢复点目标），良好的同步策略要求 RPO 时间尽量接近、等于故障发生时间，达到业务数据的最小差异或者零差异，同时也要做到不会因为频繁的数据同步带来大量的网络传输，占用服务器及网络带宽资源，从而影响到了业务生产的使用。

CMS 系统的业务数据包括了应用程序、产生的业务数据文件（html 页面、图片、视频文件等）和数据库。

3.2.1 冗余的部署结构

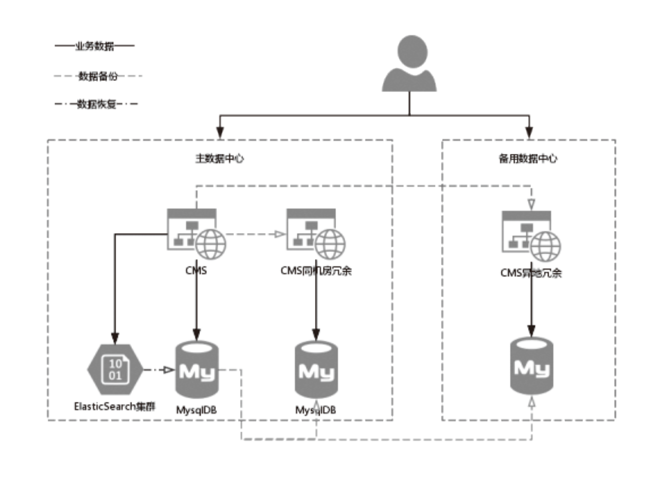


图 1

如图 1 所示，CMS 系统部署在两个数据中心，实现了数据中心级别的冗余。网络层面，主机到交换机，交换机到上层出口设备均使用了冗余线路连接。

主数据中心，部署了两套 CMS 系统用于实现主机级别的冗余。

两套 Web 系统部署在独立的物理主机中，连接各自的数据库，两套系统均可以独立工作，之间没有任何业务的依赖和关联，确保了当一套系统出现故障时，可以迅速的切换至备用系统

3.2.2Binlog 备份同步方案

数据的备份，同步使用操作系统的计划任务程序实现，根据配置好的计划任务配合 rsync 命令来实现各种同步策略。

详细同步策略说明如下：

应用程序的同步策略，由于应用程序开发改造版本迭代没有时间规律，为即刻使新版本程序生效，且不希望使用低效的 crontab 定时机制，造成高频度的资源消耗，我们采用了 Linux 的 signal 信号机制实现了这个功能，即新版本上传至指定目录后，系统自动侦测到文件有改动后，马上启动同步脚本将新的程序文件同步到同网段和异网段的备份主机中，快捷而高效，对系统的资源消耗较小。

数据库全备份，数据库的全备份在系统重构环节是必不可少的，但是数据库全备份会对生产系统造成较大的影响，由于数据保有量极大，备份过程会造成 30 分钟以上的系统“伪瘫痪”，无法承载高负荷的系统运行。所以，我们严格控制全备份的频率，每天仅在凌晨 2 点进行一次数据库的全备份脚本操作。

数据库 binlog 文件同步，binlog 日志记录着数据库所有 DDL 和 DML 语句，因此包含了数据结构和内容的变化，将 binlog 进行备份的意义在于，通过恢复 binlog 不但能够保证数据库的内容最终一致性，重要的是可以记录数据变化的过程，一旦出现数据丢失、误操作等情况，可以根据冷备份和 binlog 将数据完全还原，且 Binlog 日志收集传输与数据库全备份的更大区别是，不会使系统运行缓慢，对业务影响很小。根据业务情况，我们设定 binglog 增量文件同步周期为 5 分钟。

3.2.3 数据库内容恢复

应急处理方案是备份恢复的重中之重，只有备份计划，如果应急处理方案不合理，将直接影响整体容灾方案的效果，无法满足业务连续性。CMS 系统分别针对不同的故障制定了对应的应急处理方案。

（1）当系统主节点出现宕机或者服务不可用时，zabbix 监控和应用监控可以及时发现故障并通过短信，微信，邮件的方式告知运维人员。

（2）运维人员确认故障严重程度，如果短时（15 分钟以内）可以在本机进行故障的修复，则进行故障的修复。如果判断为严重的系统或者硬件故障，无法短时在本机进行修复则启动应急切换流程。

（3）根据应急切换流程将系统切换至备用系统中。

（4）在备用系统中验证业务有效性

（5）备用系统上线提供服务

（6）业务系统恢复使用后，进行主节点系统的修复工作

修复完成后验证主节点业务及数据情况

3.2.4 系统切换方法

在前面介绍的整体应急响应流程中，最关键的环节就是系统切换环节。备份策略的制定，都是为了在故障发生时能够有效的进行系统切换。

（1）Web 系统的切换

故障发生在 CMS 系统所在主机上，该网段其它设备状况良好，则将同网段的备份系统启用。之前的备份策略提到，备份系统中，应用程序以及应用程序所产生的数据

文件都是实时同步，所以备用系统的 web 服务是实时可用的，web 系统的切换方法为对调主备服务器的 IP 地址，这样做的好处在于，只在 OSI 模型中的第三层 IP 层做了修改，对于上层应用没有任何变化，对于 DNS 解析以及其它相关联系统的调用没有任何变更的感知，无需其它系统做任何调整既可以使备用的 web 系统上线使用。

如遇到数据中心级故障（如出口网络设备故障、运营商线路故障、电力中断等），web 系统需要切换到异地机房的备份 web 系统中，此时需要修改系统的域名解析，并启用备份系统，由于是内部系统，并使用了内部的 DNS 服务器，因此 DNS 生效时间可较短，根据目前 DNS 服务器设定的 TTL 时间为 600 秒。

（2）数据的切换及校验

前文备份策略中提到，备份系统中的数据库文件采用周期同步策略进行冷备，同机房和异地备份机房同步周期均为 1 天，如果只采用冷备数据作为备用系统的数据源，从 RPO（恢复点目标）的指标来看数据差异过大，就相同机房而言，业务数据理论上最大会出现 24 小时的数据滞后，这对于 CMS 系统而言是不能接受的。因此需要进一步进行差异数据的恢复。

这里分别从数据库视角和业务视角进行恢复和验证，首先数据库恢复程序对 mysql 数据库完成细粒度的恢复，保证数据库的条目完整，之后数据验证程序会根据 ElasticSearch 中的具体业务数据（稿件，专题）等对数据的恢复的效果及完整性进行自动校验。

3.3 数据恢复

数据恢复分为两个步骤，数据库全库恢复及增量添加。

全库恢复每天进行一次，在每日凌晨 2 点进行数据库备份后，将运行文件同步脚本至备份主机，并进行数据库的装载和修复工作，使之达到可用状态。

增量添加的基本流程：主系统产生的 Binlog 日志与备份系统进行 5 分钟一次的实时同步。当主系统发生故障后，备份系统会导入当天凌晨 2 点以后的多个 binlog 文件增量数据，将差异数据进行恢复至最近的状态。根据 binlog 日志恢复，确定当前备份系统 mysql 数据的最后写入时间点，然后使用 mysql 自带的 mysqlbinlog 工具导出为 sql 文件，使用此 sql 文件导入 mysql 数据库完成数据的恢复。

3.4 数据自动校验与回填

ElasticSearch 数据库集群，在网站的发布系统体系中，可承担多重作用：

作为全文检索数据库。由于 ElasticSearch(简称 ES）构建于著名的开源全文数据库 Lucene，承担网站全文检索是非常自然的功能需求。

作为动态内容的展示平台。大型资讯类网站新闻展示页、首页等页面多以文件静态方式呈现，而类似列表页、更多页这样的页面更适合采用动态的展示方式来呈现，当用户浏览栏目页、内容更多页时，可使用 ES 来进行动态展示。

作为数据自动校验工具。由于 ES 库中保留着编辑所发的任何一条稿件及其主要版本，利用其这一特性，可用于在系统故障迁移完成后，自动对迁移后的数据库完整性进行校验，防止稿件丢失，这个对比过程可通过预先编写好的脚本完成。此处主要详细叙述 ES 作为数据校验工具时的业务流程。

发布系统生成内容数据时，一边将内容写到 mysql 数据库，同时也将关键的内容数据写入到 ES 集群中，ES 集群采用冗余结构部署，具有较高的可用性，单机故障不影响集群的健康使用。

发布系统数据库恢复完成后，启用数据自动检查程序，检查程序会对故障时间点到当前的业务数据逐条进行检查，验证 mysql 数据库是否包含此条目，不同于 mysql 校验的是，这里不是按照 mysql 的数据结构进行检验，而是按照实际业务发生的数据进行检验，也就是会从 ES 中按时间顺序查出每条文稿信息，在 mysql 相关表中进行检索，如果发现 mysql 数据库中缺少相关记录，则将缺少的内容补充到 Mysql 数据库的记录中。

3.5 备份方案实际演练及评估

备份方案的可行性需要真实的演练验证，针对本系统，分别模拟 CMS 系统主节点单机故障和数据中心网络故障。

（1）演练方法，对于系统单机故障，采用关闭主机的方法；数据中心网络故障的模拟则直接修改 web 系统的 dns 解析记录。

（2）演练效果

序号	故障类型	CMS 容灾方案	传统容灾方案
1	Web 系统站点故障	RPO=5 分钟，RTO=5 分钟	RPO=5 分钟，RTO=5 分钟
2	数据库单机故障	RPO=10 分钟，RTO=5 分钟	RPO=10 分钟，RTO<4 小时
3	数据中心级故障	RPO=15 分钟，RTO=5	RPO=15 分钟，RTO< 一天

表 1

如表 1，记录了 CMS 容灾方案在实际演练中的 RPO 和 RTO 时间，我们可以看出，遇到 web 系统的故障，CMS 系统容灾和传统的容灾方案因为都采用了文件实时同步方式，所以恢复时间上没有差距，但如果发生数据库系统的故障或者数据中心级别的故障，CMS 系统的 RPO 时间短于传统的容灾方案。RPO 时间则明显短于传统容灾方案，对于数据可用性带来本质的提升。也就是说，当前 CMS 系统的容灾方案可以做到当遇到严重的系统故障时，可以在 10 分钟内恢复系统使用，并且 20 分钟内可以将数据恢复至系统故障前的 5 分钟内。

（作者单位：中国新闻社）